BMJ Open

Reporting and interpretation of results from clinical trials that did not claim a treatment difference

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-024785
Article Type:	Research
Date Submitted by the Author:	03-Jul-2018
Complete List of Authors:	Gates, Simon; University of Birmingham, Cancer Research UK Clinical Trials Unit; University of Warwick, Clinical Trials Unit Ealing, Elizabeth; University of Warwick, Warwick Clinical Trials Unit
Keywords:	Clinical trials < THERAPEUTICS, reporting, STATISTICS & RESEARCH METHODS

Reporting and interpretation of results from clinical trials that did not claim a treatment difference

Simon Gates [1], Elizabeth Ealing [2]

[1] Professor of Clinical Trials, Warwick Clinical Trials Unit, University of Warwick, Coventry CV4 7AL, UK

[2] Student, Warwick Clinical Trials Unit, University of Warwick, Coventry CV4 7AL, UK

[1] current address: Cancer Research UK Clinical Trials Unit, University of Birmingham, Birmingham B15 2TT

[1] Corresponding author. Email address s.gates@bham.ac.uk

Abstract

Objectives: To describe and summarise the reporting of ”non-significant” results in clinical trials, and to estimate how commonly clinical trial reports make an erroneous claim of no treatment difference based on a non-statistically significant result.

Design: Retrospective survey.

Setting: Four high impact factor general medical journals, published between June 2016 and June 2017.

Participants: Reports of randomised controlled trials that did not find a difference between the interventions they compared.

Interventions: No intervention.

Primary and secondary outcome measures: We used a 10-category classification for the text describing results for the primary outcome or outcomes, in the Results and Conclusions sections of the Abstract of each paper. Proportion of papers making claims that were not justified by the results.

Results: Eighty-five trial reports were included, reporting 111 treatment comparisons. The majority of papers (55%) concluded that there was no treatment benefit. The other common approaches were to state that there was no significant benefit (12.6%) or no significant difference (11.7%).

Conclusions: Despite decades of warnings, the error of concluding a lack of treatment benefit from a non-statistically significant result remains common.

Keywords: clinical trial, reporting, statistics

1. Strengths and limitations of this study

- We surveyed every issue of four journals for a recent 12-month period, hence the results are comprehensive and up to date.
- Restriction to four high-impact general journals means that we cannot draw any conclusions about specialised or lower impact publications.
- Our classification system was developed by the authors and is not a validated tool.

2. Introduction

Reports of randomised controlled trials (RCTs) usually attempt to draw conclusions about treatment effectiveness from their statistical analysis. It is common for results that pass a threshold for statistical significance (usually p<0.05) to be interpreted as indicating a real and clinically important effect, whereas non-significance is often taken to mean that there is no difference between the treatments, or that the intervention is not effective. As has been pointed out many times, this is an erroneous conclusion.[1][2] Failure to reach a conventional threshold for ”statistical significance” does not mean that it is safe to conclude that there is no difference. In a trial with 80% power, non-significance is expected 20% of the time, if the true treatment effect as large as expected in the trial planning, and lower power, which is commonly caused by a smaller number of recruits or a true treatment effect that is smaller than was assumed, gives a higher probability of non-significance. Misinterpretation of non-significant results in clinical trials may be particularly damaging, because trials are regarded as the highest standard of evidence, and their results often determine clinical guidelines and practice. Erroneous conclusions of ineffectiveness may result in non-adoption or abandonment of treatments that could actually be beneficial, and the existence of an apparently “definitive” trial that concluded ineffectiveness is likely to discourage further research. This problem was identified over twenty years ago[3] (“absence of evidence is not evidence of absence”), and several subsequent studies have documented its persistence.[4][5]

The motivation for this study was our observation that, despite these warnings, poor interpretations of non-conclusive trial results remain common, even in the most prestigious journals. We examined how results were described when the data did not show a statistically significant difference between the treatment arms, in the Abstracts of recent reports of randomised controlled trials (RCTs) published in four leading general medical journals.

3. Methods

We hand searched issues of four journals (New England Journal of Medicine, Journal of the American Medical Association, The Lancet and British Medical Journal) published between June 2016 and June 2017. Papers were included if they were primary reports of RCTs that had non-significant results for their primary outcome. We excluded non-inferiority, equivalence, and single armed trials, as they have different reporting issues. We included multiple-armed trials, and trials with multiple primary comparisons, if no treatment difference was claimed for any of them.

We extracted information from the abstract of each report on the description (from the Results or Findings section) and interpretation (from the Conclusions or Interpretation section) of the trials results for the primary outcome or outcomes. We concentrated on the abstracts because these are the most frequently viewed parts of papers, so conclusions expressed here will have the most impact. We classified the descriptions into ten categories (Box 1). We also recorded whether confidence intervals and p-values were presented, and whether they were referred to in conclusions.

Data were extracted by both authors independently and discrepancies resolved by discussion.

4. Results

We identified 85 eligible trial reports, reporting 111 treatment comparisons. Three journals published most of the studies (JAMA: 26, Lancet 26, NEJM 28, BMJ 5). The majority of studies used a p-value of 0.05 as the cutoff for statistical significance; two studies used lower threshold values (0.04 and 0.01), to correct for multiple comparisons. Significance tests were presented for 87/111 (78.3%) comparisons (72/85 papers (84.7%)), and confidence intervals for 88/111 (79.3%) comparisons (71/85 papers (83.5%)); all were 95% confidence intervals, except for the two studies that used different significance levels.

In the results section (Figure 1), the commonest reporting style was to present the point estimate and confidence interval, without any interpretation (55/111; 49.5%), with substantial numbers also referring to lack of statistical significance (34/111; 30.6%) or stating that there was no difference (8/111; 7.2%) or no improvement (7/111; 6.3%).

In the conclusions (Figure 2), a substantial majority of comparisons were classified as stating that there was no treatment benefit (61/111; 55.0%). The main alternative approach was to re-state the lack of a statistically significant difference (13/111; 11.7%) or statistically significant benefit (14/111; 12.6%). Only 4/111 (3.6%) comparisons (3 studies) explicitly referred to the confidence interval or uncertainty around the treatment effect estimate when drawing conclusions.

5. Discussion

The majority of the trials concluded, based on non statistically significant results, that the treatment being evaluated did not improve outcomes. Several types of result could give rise to such statements. One possible meaning is that the results demonstrated that improvement in outcomes was unlikely; the treatment effect was estimated precisely enough to make clinically important benefit unlikely. A second possibility is that the results were inconclusive; substantial uncertainty remained and neither benefit nor harm could be excluded. Yet another possibility is that the trial suggested benefit, but not convincingly enough to allow a conclusion of superiority. Consideration of the uncertainty around the treatment effect estimate would help to distinguish between these possibilities, but only 3.6% of comparisons referred to the uncertainty when drawing conclusions. The language used was often open to multiple interpretations. A statement that an intervention "did not improve" an outcome could be understood either as meaning that the study demonstrated that there was no improvement, or that improvement was not demonstrated, but remained possible. It seems particularly problematic to conclude lack of treatment benefit when there is substantial uncertainty about the direction and size of the treatment effect, or when the results are strongly in one direction. For example, one trial concluded that the incidence of the outcome was "not reduced" by the intervention, based on a risk ratio of 1.13 (95% confidence interval 0.63, 2.00),[6] and another concluded that the intervention was "not found to be superior" where the hazard ratio was 0.89 and the 95% confidence interval 0.78 to 1.01.[7]

A further 27% of comparisons qualified their conclusion of lack of treatment benefit by referring to statistical significance. This makes more explicit that a threshold p-value was used to make the judgement, but again, it is unclear exactly what meaning is intended; is it intended to be synonymous with "no difference," or to leave open the possibility that a difference may exist but has not been found? It is unclear whether referring to statistical significance helps interpretation, as there is substantial empirical evidence that this concept is often misinterpreted by the public[8], academic researchers[9], and statisticians.[10]

All of the trials in our sample used traditional frequentist statistical methods to draw conclusions. Although this is the dominant statistical methodology in clinical trials, there are many problems in the understanding and interpretation of p-values, significance tests,[11][12][1] and confidence intervals,[13][14] which have recently received substantial publicity, in the wake of publication of the American Statistical Association's guidance on p-values and significance testing.[2] One important issue is that use of a threshold for "significance" creates a binary classification of results, which is interpreted as indicating treatments that "work" and "don't work" (or "positive" and "negative" trials, or "effective" and "ineffective" treatments).[15][16][17] In reality there is no such sharp dividing line between treatments that work and do not work, and significance tests simply impose an arbitrary criterion. The persistence of dichotomisation of results may be due to an unrealistic expectation that trials will provide certainty in their conclusions and treatment recommendations. Sometimes trials will reduce our uncertainty sufficiently that the best clinical course of action is clear, but often they will not.

How can we do better? One straightforward way is to be more careful about the language that is used to describe results and draw conclusions, and ensure that written descriptions match the numerical results. We should avoid language that is ambiguous or open to misinterpretation, for example only describing treatments as ineffective if we can be sufficiently sure that the treatment does not have clinically important effects. We should also pay more attention to uncertainty, and consider what possible values of the unknown underlying treatment effect could have given rise to the data that were observed. Often, there will be a wide range of true treatment effects that could plausibly have led to the observed data. We should not expect every trial to lead to a clear treatment recommendation, but be honest about the degree to which a study is able to reduce our uncertainty. Confidence intervals were originally promoted for trial reporting to encourage this sort of interpretation, and to avoid the false certainty provided by significance tests.[18][19] But even though most trials now present them, they are rarely considered in the conclusions,[20][21] and are often used simply as an alternative way to perform significance tests (if the null value is outside the 95% confidence interval, then the p-value will be less than 0.05).

A more radical solution is to change the statistical approach that we use. One fundamental problem with traditional frequentist statistical methods is that they do not provide the results that clinicians, policy makers and patients actually want. We want to know what are the most plausible values of the treatment effect, given the observed data. Significance tests actually do the reverse; they calculate probabilities of the data (or more extreme data), assuming a specific value (usually zero) of the treatment effect. We need to use Bayesian statistical methods to get the probabilities that we want. The output from a Bayesian analysis is a probability distribution giving the probability of all possible values of the treatment effect, taking into account the the trial's data, and if desired, external information as well. We can use this distribution (the posterior probability distribution) to calculate relevant and informative results, such as the probability of a benefit exceeding a threshold for clinical importance, the probability of the treatment effect being within a range of clinical equivalence, or the range of treatment effects with 95% probability (or 50%, or any other value) of including the true value. Frank Harrell's blog gives examples of the sorts of informative statements that can be made from Bayesian results (http://www.fharrell.com/post/bayes-freq-stmts/). In particular, there is no need to reduce results to an artificial dichotomy. Because Bayesian methods deal directly with the probabilities of the possible values of the treatment effect, they are much better aligned with the underlying scientific questions.

6. Conclusions

Despite being identified over 30 years ago, and the publication of regular warnings, the "absence of evidence is not evidence of absence" error is still frequently committed in reports of RCTs in high-impact journals. Dichotomisation of results by significance tests encourages this misinterpretation and the unrealistic expectation that RCTs will always be able to give conclusive clinical results. Interpretation of results should pay more attention to uncertainty and the range of treatment effects that could plausibly have given rise to the observed data, and a switch to Bayesian statistical methods would facilitate this.

7. Funding and Ethics

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. Ethics Committee approval was not required.

8. Copyright Statement

9. Competing Interests

Both authors have completed the Unified Competing Interest form and declare no support from any organisation, no financial relationships with any organisations that might have an interest in this work, and no other relationships or activities that could appear to have influenced the submitted work.

10. Contributors

Simon Gates designed the study, assisted with data extraction, performed the analysis and drafted the manuscript. Elizabeth Ealing collected the data, assisted with analysis, and revised the manuscript. Simon Gates is the guarantor.

11. Transparency statement

The lead author affirms that this manuscript is an honest, accurate, and transparent account of the study being reported.

12. Patient and public involvement

There was no patient and public involvement in this study.

13. Data Sharing Statement

The data from this study are available from the first author.

14. References

[1] Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., et al. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. European journal of epidemiology 2016;31(4):337–350.

[2] Wasserstein, R., Lazar, N.. The asas statement on p-values: context, process, and purpose. The American Statistician 2016;70(2):129–133.

[3] Altman, D.G., Bland, J.M.. Statistics notes: Absence of evidence is not evidence of absence. BMJ 1995;311(7003):485.

[4] Alderson, P., Chalmers, I.. Survey of claims of no effect in abstracts of cochrane reviews. BMJ 2003;326(7387):475.

[5] Greenland, S.. Null misinterpretation in statistical testing and its impact on health risk assessment. Preventive medicine 2011;53(4):225–228.

[6] Thomusch, O., Wiesener, M., Opgenoorth, M., Pascher, A., Woitas, R.P., Witzke, O., et al. Rabbit-atg or basiliximab induction for rapid steroid withdrawal after renal transplantation (harmony): an open-label, multicentre, randomised controlled trial. The Lancet 2016;388(10063):3006–3016.

[7] Johnston, S.C., Amarenco, P., Albers, G.W., Denison, H., Easton, J.D., Evans, S.R., et al. Ticagrelor versus aspirin in acute stroke or transient ischemic attack. New England Journal of Medicine 2016;375(1):35–43.

[8] Tromovitch, P.. The lay public's misinterpretation of the meaning of'significant': A call for simple yet significant changes in scientific reporting. Journal of Research Practice 2015;11(1):1.

[9] Haller, H., Krauss, S.. Misinterpretations of significance: A problem students share with their teachers. Methods of Psychological Research 2002;7(1):1–20.

[10] McShane, B.B., Gal, D.. Blinding us to the obvious? the effect of statistical training on the evaluation of evidence. Management Science 2015;62(6):1707–1718.

[11] Goodman, S.. A dirty dozen: twelve p-value misconceptions. In: Seminars in hematology; vol. 45. Elsevier; 2008, p. 135–140.

[12] Goodman, S.N.. Toward evidence-based medical statistics. 1: The p value fallacy. Annals of internal medicine 1999;130(12):995–1004.

[13] Hoekstra, R., Morey, R.D., Rouder, J.N., Wagenmakers, E.J.. Robust misinterpretation of confidence intervals. Psychonomic bulletin & review 2014;21(5):1157–1164.

[14] Morey, R.D., Hoekstra, R., Rouder, J.N., Lee, M.D., Wagenmakers, E.J.. The fallacy of placing confidence in confidence intervals. Psychonomic bulletin & review 2016;23(1):103–123.

[15] McShane, B.B., Gal, D.. Statistical significance and the dichotomization of evidence. Journal of the American Statistical Association 2017;112(519):885–895.

[16] Senn, S.. Dichotomania: an obsessive compulsive disorder that is badly affecting the quality of analysis of pharmaceutical trials. Proceedings of the International Statistical Institute, 55th Session, Sydney 2005;.

[17] Gelman, A., Stern, H.. The difference between significant and not significant is not itself statistically significant. The American Statistician 2006;60(4):328–331.

[18] Moher, D., Hopewell, S., Schulz, K.F., Montori, V., Gøtzsche, P.C., Devereaux, P., et al. Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869.

[19] Gardner, M.J., Altman, D.G.. Confidence intervals rather than p values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed) 1986;292(6522):746–750.

[20] Fidler, F., Thomason, N., Cumming, G., Finch, S., Leeman, J.. Editors can lead researchers to confidence intervals, but can't make them

think: Statistical reform lessons from medicine. Psychological Science 2004;15(2):119–126.

[21] Gewandtner, J.S., McDermott, M.P., Kitt, R.A., Chaudari, J., Koch, J.G., Evans, S.R., et al. Interpretation of cis in clinical trials with non-significant results: systematic review and recommendations. BMJ open 2017;7(7):e017288.

Box 1. Classification of reporting of results

1. No treatment difference, including did not differ, no difference, no effect, no change,

2. No treatment difference, qualified by reference to statistical significance; including no significant difference, not statistically different, not statistically significant, no significant effect."

3. No treatment difference, qualified by something other than statistical significance, including no substantial difference, no clinically relevant difference.

4. No treatment benefit, including did not result in increase/decrease/improvement, was not superior, did not increase/decrease/improve, did not prevent.

5. No treatment benefit, qualified by reference to statistical significance, including not significantly better/worse, did not significantly increase/decrease, not statistically increased/decreased.

6. No treatment benefit, qualified by reference to something other than statistical significance, including not substantially increased/decreased.

7. Lack of evidence for a difference.

8. The treatments compared were similar.

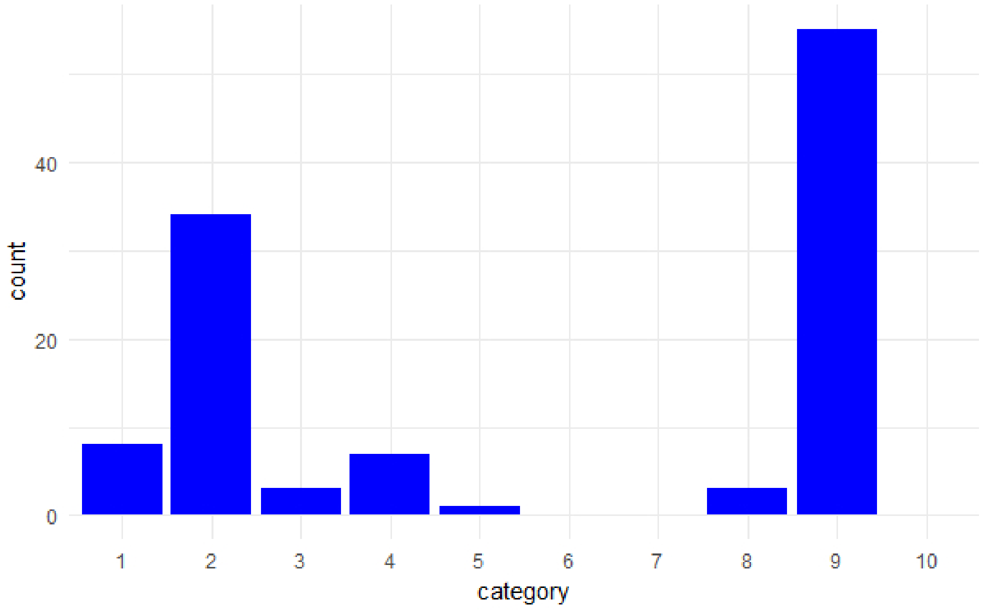
9. Statement of the results, without any claim about the size or direction of effect.

10. Clinical recommendation without interpretation of results.

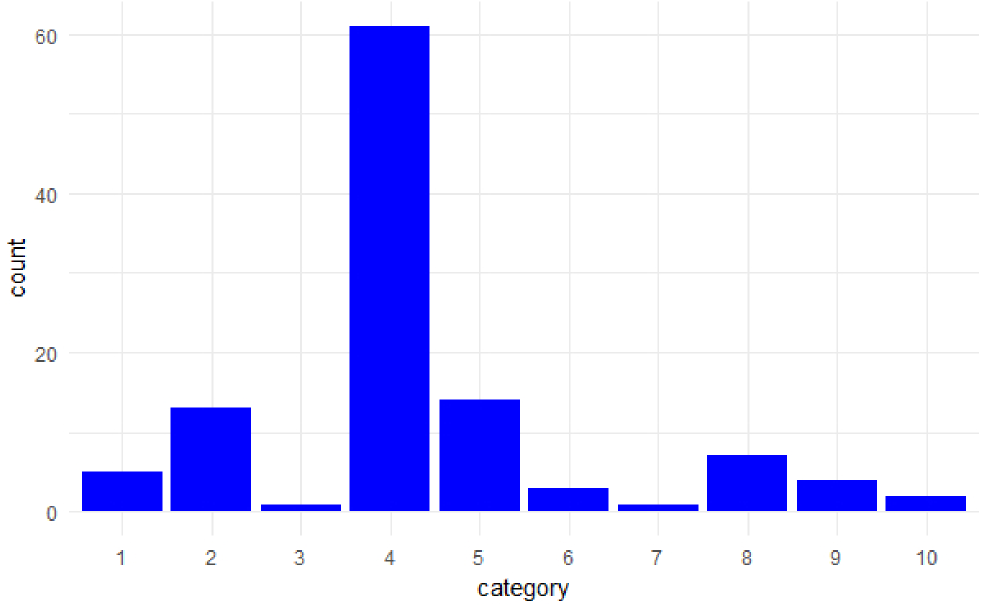
13. Figure legends

Figure 1: Frequencies of different types of description of results in Results section of abstracts. Categories (described fully in Box 1): 1. no difference; 2. no statistically significant difference; 3. no substantial or clinically important difference; 4. no improvement or no treatment benefit; 5. no significant improvement; 6. no substantial improvement; 7. lack of evidence for a difference; 8. treatments were similar; 9. statement of results; 10. clinical recommendation

Figure 2: Frequencies of different types of description of results in Conclusions section of abstracts. Categories (described fully in Box 1): Categories (described fully in Box 1): 1. no difference; 2. no statistically significant difference; 3. no substantial or clinically important difference; 4. no improvement or no treatment benefit; 5. no significant improvement; 6. no substantial improvement; 7. lack of evidence for a difference; 8. treatments were similar; 9. statement of results; 10. clinical recommendation



Frequencies of different types of description of results in Results section of abstracts. Categories (described fully in Box 1): 1. no difference; 2. no statistically sig- nificant difference; 3. no substantial or clinically important difference; 4. no improvement or no treatment benefit; 5. no significant improvement; 6. no substantial improvement; 7. lack of evidence for a difference; 8. treatments were similar; 9. statement of results; 10. clinical recommendation



Frequencies of different types of description of results in Conclusions section of abstracts. Categories (described fully in Box 1): Categories (described fully in Box 1): 1. no difference; 2. no statistically significant difference; 3. no substantial or clinically impor- tant difference; 4. no improvement or no treatment benefit; 5. no significant improvement;
6. no substantial improvement; 7. lack of evidence for a difference; 8. treatments were similar; 9. statement of results; 10. clinical recommendation

BMJ Open

Reporting and interpretation of results from clinical trials that did not claim a treatment difference; survey of four general medical journals

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-024785.R1
Article Type:	Research
Date Submitted by the Author:	16-Jan-2019
Complete List of Authors:	Gates, Simon; University of Birmingham, Cancer Research UK Clinical Trials Unit; University of Warwick, Clinical Trials Unit Ealing, Elizabeth; University of Warwick, Warwick Clinical Trials Unit
<b>Primary Subject Heading</b>:	Research methods
Secondary Subject Heading:	Research methods
Keywords:	Clinical trials < THERAPEUTICS, reporting, STATISTICS & RESEARCH METHODS

Reporting and interpretation of results from clinical trials that did not claim a treatment difference; survey of four general medical journals

Simon Gates [1], Elizabeth Ealing [2]

[1] Professor of Clinical Trials, Warwick Clinical Trials Unit, University of Warwick, Coventry CV4 7AL, UK

[2] Student, Warwick Clinical Trials Unit, University of Warwick, Coventry CV4 7AL, UK

[1] current address: Cancer Research UK Clinical Trials Unit, University of Birmingham, Birmingham B15 2TT

[1] Corresponding author. Email address s.gates@bham.ac.uk

Abstract

Objectives: To describe and summarise the reporting of “non-significant” results in randomized controlled trials (RCTs), and to estimate how commonly trial reports make erroneous claims of no treatment difference based on a non-statistically significant result.

Design: Retrospective survey of published RCTs.

Setting: Four high impact factor general medical journals, published between June 2016 and June 2017.

Participants: Reports of randomised controlled trials that did not find a difference between the interventions they compared.

Interventions: Not an interventional study.

Primary and secondary outcome measures: We recorded the way each trial’s results for its primary outcome or outcomes were described in the Results and Conclusions sections of the Abstract, using a 10-category classification. We estimated the proportion of papers that made claims that were not justified by the results, or were open to multiple interpretations.

Results: Eighty-five trial reports were included, reporting 111 treatment comparisons. The majority of papers made unjustified or confusing statements. In the Results section of abstracts, for 55/111 comparisons (49.5%) the study’s results were re-stated, without interpretation, and 34/111 (30.6%) stated that there was not a statistically significant difference. In the conclusions, 61/111 treatment comparisons (55%) stated that there was no treatment benefit, 14/111 (12.6%) that there was no significant benefit, and 13/111 (11.7%) that there was no significant difference.

Conclusions: Despite decades of warnings, the error of concluding a lack of treatment benefit from a non-statistically significant result remains common.

Keywords: clinical trial, reporting, interpretation, statistics

Strengths and limitations of this study

- We surveyed every issue of four journals for a recent 12-month period, hence the results are comprehensive and up to date.
- This was not a systematic review, but was restricted to four high-impact general journals. This means that we cannot draw any conclusions about other publications.
- Our classification system was developed by the authors and is not a validated tool.
- We only looked at reporting in abstracts; in the main text of papers authors may have made different and more accurate statements.

Introduction

Reports of randomised controlled trials (RCTs) usually attempt to draw conclusions about treatment effectiveness from their statistical analysis. It is common for results that pass a threshold for statistical significance, usually a p-value of less than 0.05, to be interpreted as indicating a real and clinically important effect. “Non-significance” (p>0.05) is often taken to mean that there is no difference between the treatments, or that the intervention is not effective. As has been pointed out many times, this is an erroneous conclusion.[1][2] Failure to reach a conventional threshold for ”statistical significance” does not mean that it is safe to conclude that there is no difference. Every statistical test has a Type II error rate, which is the probability of obtaining a non-significant result, if the null hypothesis is false i.e. that there really is a difference. Trials are often designed with a 20% Type II error rate (80% power), for a true treatment effect of a specified size. With such a design, even if the true treatment effect is exactly as assumed (and designs often assume unrealistically large treatment effects), non-significance would be expected 20% of the time, and a conclusion of no difference would then be wrong. Moreover, common issues such as fewer recruits than expected, more variability, or a lower incidence of outcomes, will reduce power, and make non-significant results more likely, even if in reality there is a real and important treatment effect. There is no way of discriminating between non-significant results that derive from chance or lack of power, and those that derive from a true lack of treatment benefit, except by more research.

Misinterpretation of non-significant results in clinical trials may be particularly damaging, because trials provide high-quality evidence, and their results often determine clinical guidelines and practice. Erroneous conclusions of ineffectiveness may result in non-adoption or abandonment of treatments that could actually be beneficial, and the existence of an apparently "definitive" trial that concluded ineffectiveness is likely to discourage further research. This problem was identified over twenty years ago[3] ("absence of evidence is not evidence of absence"), and subsequent studies have documented its persistence.[4][5]

The motivation for this study was our observation that, despite these warnings, poor interpretations of non-conclusive trial results remain common, even in the most prestigious journals. Many trials where the main results are not statistically significant conclude that there is no difference between the treatments, the intervention did not improve outcomes, or that it was not effective, none of which is a justified interpretation.

We examined how results were described in the Abstracts of recent reports of RCTs where the primary outcome did not show a statistically significant difference between the treatment arms, published in four leading general medical journals

Methods

We hand searched issues of four journals (New England Journal of Medicine (NEJM), Journal of the American Medical Association (JAMA), The Lancet and British Medical Journal (BMJ)) published between June 2016 and June 2017. Papers were included if they were primary reports of RCTs that had non-significant results for their primary outcome. We excluded non-inferiority, equivalence, and single armed trials, as they have different reporting issues. We included multiple-armed trials, and trials with multiple primary comparisons, if no treatment difference was claimed for any of them.

We extracted information from the abstract of each report on the description (from the Results or Findings section) and interpretation (from the Conclusions or Interpretation section) of the trials results for the primary outcome or outcomes. We concentrated on the abstracts because these are the most frequently viewed parts of papers, so conclusions expressed here will have the most impact. We classified the descriptions into ten categories (Table 1). The classification was developed at the start of the project, by reviewing trial reports from the same journals that were published in January to May 2016, the period immediately before our study's eligibility window. The classification made a distinction between reporting that claimed a lack of directional effect (e.g. "no improvement") and reporting that did not include any directional information (e.g. "no difference"), as well as whether the claim was qualified by reference to statistical significance (e.g. "no significant difference") or something else (e.g. "no substantial difference"). We created additional categories during the study, for reports that used methods that did not fit into any of the predetermined categories; for example, statements such as "there was a lack of evidence for a difference," or "treatments were simila.," We also recorded whether confidence intervals and p-values were presented, and whether the confidence interval, or uncertainty more generally, were referred to in the conclusions.

Data were extracted by both authors independently and discrepancies resolved by discussion.

Results

We identified 85 eligible trial reports, reporting 111 treatment comparisons. Three journals published most of the studies (JAMA 26, Lancet 26, NEJM 28, BMJ 5). The majority of studies used a p-value of 0.05 as the cutoff for statistical significance; two studies used lower threshold values (0.04 and 0.01), to correct for multiple comparisons. Significance tests were presented for 87/111 (78.3%) comparisons (72/85 papers (84.7%)), and confidence intervals for 88/111 (79.3%) comparisons (71/85 papers (83.5%)); all were 95% confidence intervals, except for the two studies that used different significance levels.

In the results section (Figure 1), the commonest reporting style was to present the point estimate and confidence interval, without any interpretation (55/111; 49.5%), with substantial numbers also referring to lack of statistical significance (34/111; 30.6%) or stating that there was no difference (8/111; 7.2%) or no improvement (7/111; 6.3%).

In the conclusions (Figure 2), a substantial majority of comparisons were classified as stating that there was no treatment benefit (61/111; 55.0%). The main alternative approach was to re-state the lack of a statistically significant difference (13/111; 11.7%) or statistically significant benefit (14/111; 12.6%).

We found that confidence intervals for the main treatment comparison were presented by only 88/111 (79.3%) studies. This was surprising, given that they have been a required part of trial reporting in the CONSORT guidelines for many years. Some trials presented confidence intervals for the difference of each randomized group from baseline, rather than the comparison between groups, or presented only p-values. These are also poor reporting practices. The proportion of trials presenting p-values was similar to the proportion that presented confidence intervals (87/111; 78.4%). We found that only 4/111 (3.6%) comparisons (3 studies) explicitly referred to the confidence interval or uncertainty around the treatment effect estimate when drawing conclusions.

Discussion

Main results

The majority of the statements (other than a simple statement of the point estimate and confidence interval), in both the Results and Conclusions sections of RCT abstracts, were not accurate or justifiable interpretations of finding a non-significant result in the analysis of the trial's primary outcome.

Statements of "no difference" or "no benefit" (categories 1 and 4), which were the most common way that results were interpreted, are problematic. Lack of statistical significance does not mean that no difference exists; this is one of the most basic misinterpretations of significance testing.[1] It is not clear from these simple statements whether the authors believed that their study had demonstrated that there was no difference or no benefit (an unjustified conclusion), or whether their view was that the evidence was insufficient to conclude that there was a treatment benefit. It seems likely that most readers would interpret a statement like "X did not improve outcome Y" as meaning that the treatment was not beneficial.

Often, studies that conclude a lack of benefit have substantial uncertainty about the direction and size of the treatment effect. For example, one trial concluded that the incidence of the outcome was "not reduced" by the intervention, based on a risk ratio of 1.13 (95% confidence interval 0.63, 2.00).[6] The confidence interval indicates that risk ratios as low as 0.63 or as high as 2.00 would be compatible with the data, so it seems unjustified to conclude that there was not benefit. The data suggest that the treatment effect can only be estimated imprecisely, so the intervention could reduce (or increase) the outcome substantially. Other trials have results that are much more suggestive of a result in one direction. An example was a trial that concluded that the intervention was "not found to be superior," with a hazard ratio of 0.89 and a 95% confidence interval of

0.78 to 1.01 [7]. The range of treatment effects compatible with the data is almost entirely in one direction. Again, the conclusion ("not found to be superior") seems inadequate; the study did suggest benefit, but not strongly enough to meet the arbitrary criterion for statistical significance.

A further 27% of comparisons qualified their conclusion of lack of treatment benefit by referring to statistical significance (categories 2 and 5). This makes more explicit that a threshold p-value was used to make the judgement, but knowing that the p-value was greater than 0.05 does not provide any useful information. It is likely that most readers would interpret significance in a way similar to the common English meaning of the word, and would take away from a "non-significant" result the impression that there was no important difference between the interventions. There is substantial empirical evidence that statistical significance is often misinterpreted by the public[8], academic researchers[9], and statisticians.[10]

Statements that there was no "substantial" difference or no "clinically important" difference (categories 3 and 6) are also difficult to interpret. It is not clear how large a difference would be regarded as "substantial," and this description may be inaccurate because non-significant differences are not necessarily small.

A statement that the interventions were "similar" (category 8), found in 3 studies' results sections and 4 studies' conclusions, is, again, vague, and often inaccurate. A non-significant result does not mean that the point estimate of the treatment effect was close to no difference, and many of the trials in this study had point estimates of the treatment effect that were far from zero. It may be reasonable to describe the results as similar if the point estimates are close, but this gives no information about the range of treatment effects that are compatible with the data.

The most reasonable way to describe non-significant results is probably that the study did not find convincing evidence against the hypothesis that that the treatment effect was zero. Only one study contained a statement that referred to lack of evidence for a difference: "We found no evidence that an intervention comprising cleaner burning biomass-fuelled cookstoves reduced the risk of pneumonia in young children in rural Malawi," [11] describing an estimated incidence rate ratio of 1.01, with 95% confidence interval 0.91 to 1.13. Hence the data were compatible with either a small increase, or a small decrease, in the risk of pneumonia.

Statistical methods

All of the trials in our sample used traditional frequentist statistical methods. Although this is the dominant statistical methodology in clinical trials, there are many problems in the understanding and interpretation of p-values, significance tests,[12][13][1] and confidence intervals,[14][15] which have recently received substantial publicity, in the wake of publication of the American Statistical Association's guidance on p-values and significance testing.[2] One important issue is the use of a threshold for "significance," creating a binary classification of results, which is interpreted as indicating treatments that "work" and "don't work" (or "positive" and "negative" trials, or "effective" and "ineffective" treatments).[16][17][18] In reality there is no such sharp dividing line between treatments that work and do not work, and significance tests simply impose an arbitrary criterion. The persistence of dichotomisation of results may be largely due to an unrealistic expectation that trials will provide certainty in their conclusions and treatment recommendations. Sometimes trials will reduce our uncertainty sufficiently that the best clinical course of action is clear, but often they will not. An argument that is often advanced in favour of dichotomization of results is that many trials seek to determine clinical practice, and a decision needs to be made about whether the intervention should be used in patient care. Two studies in our sample did indeed make recommendations for clinical practice in their conclusions. The counter-argument to this is that decisions about use of healthcare interventions should be based not on statistical significance of a single primary outcome measure, but on consideration of the overall benefits, harms and costs of the intervention, using appropriate decision modelling methodology.

Improving the language for describing results

One straightforward way to improve reporting of results is to be more careful about the language that is used to describe them and draw conclusions, and ensure that written descriptions match the numerical results. We should avoid language that is ambiguous or open to misinterpretation, for example only describing treatments as ineffective if we have a high degree of confidence that the treatment does not have clinically important effects. We should also pay more attention to uncertainty, and consider what possible values of the unknown underlying treatment effect could have given rise to the data that were observed. Often, this range will be wide. We should not expect every trial to lead to a clear treatment recommendation, but be honest about the degree to which a study is able to reduce our uncertainty. Confidence intervals were originally promoted for trial reporting to encourage this sort of interpretation, and to avoid the false certainty provided by significance tests.[19][20] But even though most trials now present them, they are rarely considered in the conclusions,[21][22] and are often used simply as an alternative way to perform significance tests (if the null value is outside the 95% confidence interval, then the p-value will be less than 0.05).

A recent online discussion [23] about language for describing frequentist trial results gave some examples of accurate statements that could be used. Three examples of statements for trials that did not find a treatment difference, from this discussion, are given below:

Example 1: “We were unable to find evidence against the hypothesis that A=B (p=0.4) with the current sample size. More data will be needed. As the statistical analysis plan specified a frequentist approach, the study did not provide evidence of similarity of A and B.”

Example 2: “Assuming the study’s experimental design and sampling scheme, the probability is 0.4 that another study would yield a test statistic for comparing two means that is more impressive that what we observed in our study, if treatment B had exactly the same true mean SBP as treatment A.”

Example 3: “Treatment B was observed in our sample of n subjects to have a 4mmHg lower mean SBP (systolic blood pressure) than treatment A with a 0.95 2-sided compatibility interval of [-13, 5], indicating a wide range of plausible true treatment effects. The degree of evidence against the null hypothesis that the treatments are interchangeable is p=0.11.”

These statements are very different from those used by most of the papers in our sample, and make much more limited claims than many real papers. However, these claims accurately reflect the conclusions that can be drawn from frequentist statistical analyses. More accurate language would help to prevent common over-interpretations, such as the belief that non-significance means that a treatment difference of zero has been established.

Improving the statistical methods

A more radical solution is to change the statistical approach that we use. One fundamental problem with traditional frequentist statistical methods is that they do not provide the results that clinicians, policy makers and patients actually want to know: what are the most plausible values of the treatment effect, given the observed data? Significance tests actually do the reverse; they calculate probabilities of the data (or more extreme data), assuming a specific null value of the treatment effect. This is a major reason why reporting frequentist results accurately is so convoluted, and why they are so difficult to understand. However, easily-interpretable probabilities of clinically relevant results can be readily obtained using Bayesian methods. The output from a Bayesian analysis is a probability distribution giving, the probability of all possible values of the treatment effect, taking into account the trial's data, and usually (via the prior), external information as well. We can use this distribution (the posterior probability distribution) to calculate relevant and informative results, such as the probability of a benefit exceeding a threshold for clinical importance, the probability of the treatment effect being within a range of clinical equivalence, or the range of treatment effects with 95% probability (or 50%, or any other value). Some examples of the sorts of informative statements that can be made from Bayesian results are given in a blog post by Frank Harrell [24]. In particular, with Bayesian methods there is no need to reduce results to an artificial dichotomy. Because Bayesian methods deal directly with the probabilities of the possible values of the treatment effect, they are much better aligned with the underlying scientific questions.

Limitations of the study

This study looked only at reporting of results in abstracts of published RCTs. We concentrated on abstracts because they are the most frequently read parts of papers, and always report the main results. They are therefore likely to be particularly important in determining readers' interpretation of the trial's results. It is possible that in other parts of the papers, reporting may have been different, and potentially more accurate. However, this is much harder to assess because results are typically reported in several different places, and often inconsistently.

We concentrated on four of the highest profile general medical journals. Obviously, RCTs are also published in a large number of other, more specialised, journals, but we cannot say whether they have the same issues of reporting as we found. Our expectation would be that, as the journals we selected are seen as some of the most prestigious publications, reporting problems would be at least as common elsewhere.

Our classification of reporting types was invented by the authors, and is not intended as a general tool for conducting this type of study. However, we feel that it is a reasonable classification that makes distinctions between the different types of reporting that we wished to identify.

Conclusions

Despite many years of warnings, inappropriate interpretation of RCT results are widespread in the most prestigious medical journals. We speculatively suggest several possible factors that may be responsible. First, authors and editors may want to present a clear message, and there is a widespread expectation that RCTs should result in clear recommendations for clinical practice. It is easier to understand a conclusion that “X did not work” than a more accurate, but more complicated, statement. Second, use of significance testing as the main analytical method provides a ready means of dichotomization of results, encouraging an over-simplified binary interpretation of interventions as effective or not effective. Third, the general difficulty of understanding frequentist results means that correct interpretation is convoluted and difficult to relate to real life.

We suggest that interpretation of results should pay more attention to uncertainty and the range of treatment effects that could plausibly have given rise to the observed data. Use of Bayesian statistical methods would facilitate this by addressing the clinical questions of interest directly.

Funding and Ethics

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. Ethics Committee approval was not required.

Copyright Statement

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, a worldwide licence to the Publish- ers and its licensees in perpetuity, in all forms, formats and media (whether known now or created in the future), to i) publish, reproduce, distribute, display and store the Contribution, ii) translate the Contribution into other languages, create adaptations, reprints, include within collections and create summaries, extracts and/or, abstracts of the Contribution, iii) create any other derivative work(s) based on the Contribution, iv) to exploit all subsidiary rights in the Contribution, v) the inclusion of electronic links from the Contribution to third party material where-ever it may be located; and, vi) licence any third party to do any or all of the above.

Competing Interests

Both authors have completed the Unified Competing Interest form and declare no support from any organisation, no financial relationships with any organisations that might have an interest in this work, and no other relationships or activities that could appear to have influenced the submitted work.

Contributors

Simon Gates designed the study, assisted with data extraction, performed the analysis and drafted the manuscript. Elizabeth Ealing collected the data, assisted with analysis, and revised the manuscript. Simon Gates is the guarantor.

Transparency statement

The lead author affirms that this manuscript is an honest, accurate, and transparent account of the study being reported.

Patient and public involvement

There was no patient and public involvement in this study.

Data Sharing Statement

The data from this study are available from Open Science Framework (https://osf.io/chsva/files/)

References

[1] Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., et al. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. European journal of epidemi- ology 2016;31(4):337–350.

[2] Wasserstein, R., Lazar, N. The asas statement on p-values: context, process, and purpose. The American Statistician 2016;70(2):129–133.

[3] Altman, D.G., Bland, J.M. Statistics notes: Absence of evidence is not evidence of absence. BMJ 1995;311(7003):485.

[4] Alderson, P., Chalmers, I. Survey of claims of no effect in abstracts of Cochrane reviews. BMJ 2003;326(7387):475.

[5] Greenland, S. Null misinterpretation in statistical testing and its impact on health risk assessment. Preventive medicine 2011;53(4):225–228.

[6] Thomusch, O., Wiesener, M., Opgenoorth, M., Pascher, A., Woitas, R.P., Witzke, O., et al. Rabbit-atg or basiliximab induction for rapid steroid withdrawal after renal transplantation (harmony): an open-label, multicentre, randomised controlled trial. The Lancet 2016;388(10063):3006–3016.

[7] Johnston, S.C., Amarenco, P., Albers, G.W., Denison, H., Easton, J.D., Evans, S.R., et al. Ticagrelor versus aspirin in acute stroke or transient ischemic attack. New England Journal of Medicine 2016;375(1):35– 43.

[8] Tromovitch, P. The lay public's misinterpretation of the meaning of 'significant': A call for simple yet significant changes in scientific reporting. Journal of Research Practice 2015;11(1):1.

[9] Haller, H., Krauss, S. Misinterpretations of significance: A problem students share with their teachers. Methods of Psychological Research 2002;7(1):1–20.

[10] McShane, B.B., Gal, D. Blinding us to the obvious? the effect of statistical training on the evaluation of evidence. Management Science 2015;62(6):1707–1718.

[11] Mortimer, K., Ndamala, C.B., Naunje, A.W., Malava, J., Katundu, C., Weston, W., et al. A cleaner burning biomass-fuelled cookstove intervention to prevent pneumonia in children under 5 years old in rural Malawi (the Cooking and Pneumonia Study): a cluster randomised controlled trial

[12] Goodman, S. A dirty dozen: twelve p-value misconceptions. In: Seminars in hematology; vol. 45. Elsevier; 2008, p. 135–140.

[13] Goodman, S.N. Toward evidence-based medical statistics. 1: The p value fallacy. Annals of internal medicine 1999;130(12):995–1004.

[14] Hoekstra, R., Morey, R.D., Rouder, J.N., Wagenmakers, E.J. Robust misinterpretation of confidence intervals. Psychonomic bulletin & review 2014;21(5):1157–1164.

[15] Morey, R.D., Hoekstra, R., Rouder, J.N., Lee, M.D., Wagenmakers, E.J. The fallacy of placing confidence in confidence intervals. Psychonomic bulletin & review 2016;23(1):103–123.

[16] McShane, B.B., Gal, D. Statistical significance and the dichotomization of evidence. Journal of the American Statistical Association 2017;112(519):885–895.

[17] Senn, S. Dichotomania: an obsessive compulsive disorder that is badly affecting the quality of analysis of pharmaceutical trials. Proceedings of the International Statistical Institute, 55th Session, Sydney 2005.

[18] Gelman, A., Stern, H. The difference between significant and not significant is not itself statistically significant. The American Statistician 2006;60(4):328–331.

[19] Moher, D., Hopewell, S., Schulz, K.F., Montori, V., Gøtzsche, P.C., Devereaux, P., et al. Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869.

[20] Gardner, M.J., Altman, D.G. Confidence intervals rather than p values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed) 1986;292(6522):746–750.

[21] Fidler, F., Thomason, N., Cumming, G., Finch, S., Leeman, J. Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. Psychological Science 2004;15(2):119–126.

[22] Gewandter, J.S., McDermott, M.P., Kitt, R.A., Chaudari, J., Koch, J.G., Evans, S.R., et al. Interpretation of cis in clinical trials with non-significant results: systematic review and recommendations. BMJ Open 2017;7(7):e017288.

[23] https://discourse.datamethods.org/t/language-for-communicating-frequentist-results-about-treatment-effects/934

[24] http://www.fharrell.com/post/bayes-freq-stmts/

Table 1. Categories of reporting of RCT results in Results and Conclusion sections of abstracts. For trials with multiple results, all were reported in the same way in all trials except one (ref); for this trial, we have included the results for the survival co-primary outcome rather than the ordinal composite outcome.

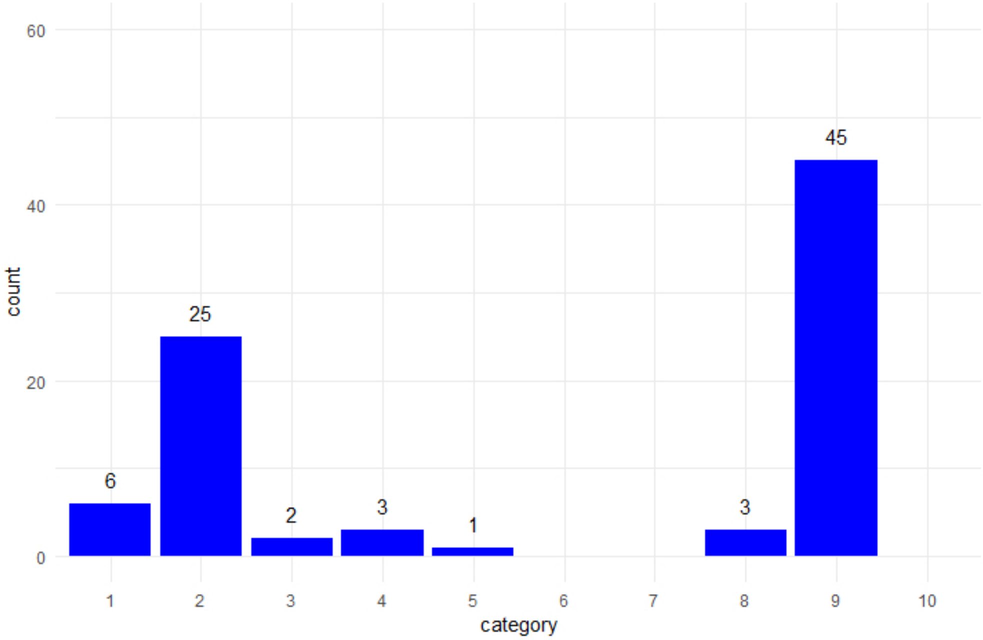
Category	**Description**	**Examples**	**Number of comparisons (%) (n=111)**		**Number of papers (%) (n=85)**	
			Results	**Conclusions**	**Results**	**Conclusions**
1	Statement of no difference between treatments	"did not differ," "no difference," "no effect," "no change."	8 (7.2)	5 (4.5)	6 (7.1)	1 (1.2)
2	Statement that there was no difference between treatment, qualified by reference to statistical significance	"no significant difference," "not statistically different," "not statistically significant," "no significant effect."	34 (30.6)	13 (11.7)	25 (29.4)	12 (14.1)
3	Statement that there was no difference between treatments, qualified by something other than statistical significance	"no substantial difference," "no clinically relevant difference."	3 (2.7)	1 (0.9)	2 (2.4)	1 (1.2)
4	Statement that the intervention was not beneficial	"did not result in increase (or decrease or improve)," "was not superior," "did not increase (or decrease or improve)," "did not prevent."	7 (6.3)	61 (55.0)	3 (3.5)	46 (54.1)
5	Statement that the intervention was not beneficial, qualified by reference to statistical significance	"not significantly better (or worse)," "did not significantly increase (or decrease)," "not statistically increased (or decreased)."	1 (0.9)	14 (12.6)	1 (1.2)	12 (14.1)
6	Statement that the intervention was not beneficial, qualified by reference to something other	"not substantially increased (or decreased)."	0 (0)	3 (2.7)	0 (0)	2 (2.4)

	than statistical significance					
7	Statement that there was a lack of evidence for a difference.	“no evidence that [intervention] reduced the risk of [outcome]”	0 (0)	1 (0.9)	0 (0)	1 (1.2)
8	Statement that the treatments compared were similar.	“yield similar outcomes” “similar risk of [outcome]” “rate of [outcome] was similar”	3 (2.7)	7 (6.3)	3 (3.5)	4 (4.7)
9	Quotation of the results, without any claim about the size or direction of effect.	Estimate and 95% confidence interval	55 (49.5)	4 (3.6)	45 (52.9)	4 (4.7)
10	Clinical recommendation, without interpretation of results.	“There is no harm in [using intervention]” “The choice between [interventions] should be made based on clinical knowledge”	0 (0)	2 (1.8)	0 (0)	2 (2.4)

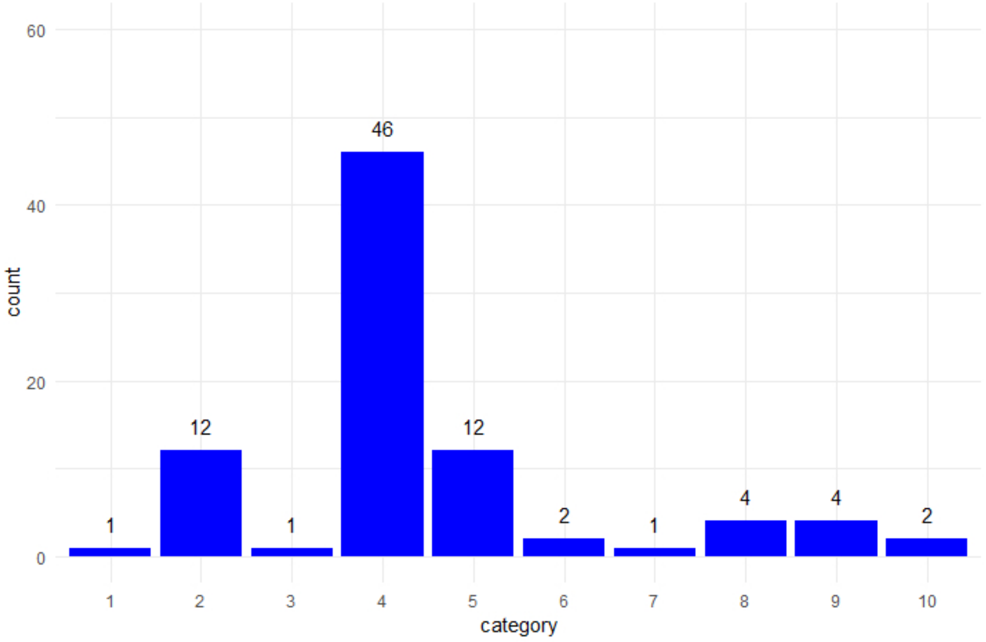
Figures

Figure 1: Frequencies of different types of description of results in Results section of abstracts. Categories (described fully in Table 1): 1. no difference; 2. no statistically significant difference; 3. no substantial or clinically important difference; 4. no improvement or no treatment benefit; 5. no significant improvement; 6. no substantial improvement; 7. lack of evidence for a difference; 8. treatments were similar; 9. statement of results; 10. clinical recommendation

Figure 2: Frequencies of different types of description of results in Conclusions section of abstracts. Categories (described fully in Table 1):): 1. no difference; 2. no statistically significant difference; 3. no substantial or clinically important difference; 4. no improvement or no treatment benefit; 5. no significant improvement; 6. no substantial improvement; 7. lack of evidence for a difference; 8. treatments were similar; 9. statement of results; 10. clinical recommendation



Frequencies of different types of description of results in Results section of abstracts. Categories (described fully in Table 1): 1. no difference; 2. no statistically significant difference; 3. no substantial or clinically important difference; 4. no improvement or no treatment benefit; 5. no significant improvement; 6. no substantial improvement; 7. lack of evidence for a difference; 8. treatments were similar; 9. statement of results; 10. clinical recommendation



Frequencies of different types of description of results in Conclusions section of abstracts. Categories (described fully in Table 1):): 1. no difference; 2. no statistically significant difference; 3. no substantial or clinically important difference; 4. no improvement or no treatment benefit; 5. no significant improvement; 6. no substantial improvement; 7. lack of evidence for a difference; 8. treatments were similar; 9. statement of results; 10. clinical recommendation

BMJ Open

Reporting and interpretation of results from clinical trials that did not claim a treatment difference; survey of four general medical journals

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-024785.R2
Article Type:	Research
Date Submitted by the Author:	03-Jun-2019
Complete List of Authors:	Gates, Simon; University of Birmingham, Cancer Research UK Clinical Trials Unit; University of Warwick, Clinical Trials Unit Ealing, Elizabeth; University of Warwick, Warwick Clinical Trials Unit
<b>Primary Subject Heading</b>:	Research methods
Secondary Subject Heading:	Research methods
Keywords:	Clinical trials < THERAPEUTICS, reporting, STATISTICS & RESEARCH METHODS

Reporting and interpretation of results from clinical trials that did not claim a treatment difference; survey of four general medical journals

Simon Gates [1], Elizabeth Ealing [2]

[1] Professor of Clinical Trials, Warwick Clinical Trials Unit, University of Warwick, Coventry CV4 7AL, UK

[2] Student, Warwick Clinical Trials Unit, University of Warwick, Coventry CV4 7AL, UK

[1] current address: Cancer Research UK Clinical Trials Unit, University of Birmingham, Birmingham B15 2TT

[1] Corresponding author. Email address s.gates@bham.ac.uk

Abstract

Objectives: To describe and summarise how the results of randomized controlled trials that did not find a significant treatment effect are reported, and to estimate how commonly trial reports make unwarranted claims.

Design: We performed a retrospective survey of published RCTs, published in four high impact factor general medical journals between June 2016 and June 2017.

Setting: Trials conducted in all settings were included.

Participants: 94 reports of randomised controlled trials that did not find a difference in their main comparison or comparisons were included.

Interventions: All interventions.

Primary and secondary outcomes: We recorded the way each trial's results for its primary outcome or outcomes were described in the Results and Conclusions sections of the Abstract, using a 10-category classification. Other outcomes were whether confidence intervals and p-values were presented for the main treatment comparisons, and whether the results and conclusions referred to measures of uncertainty. We estimated the proportion of papers that made claims that were not justified by the results, or were open to multiple interpretations.

Results: 94 trial reports (120 treatment comparisons) were included. In the Results sections, for 58/120 comparisons (48.3%) the study's results were re-stated, without interpretation, and 38/120 (31.7%) stated that there was not a statistically significant difference. In the Conclusions, 65/120 treatment comparisons (54.2%) stated that there was no treatment benefit, 14/120 (11.7%) that there was no significant benefit, and 16/120 (13.3%) that there was no significant difference. Confidence intervals and p-values were both presented by 84% of studies (79/94), but only 3/94 studies referred to uncertainty when drawing conclusions.

Conclusions: The majority of trials (54.2%) inappropriately interpreted a result that was not statistically significant as indicating no treatment benefit. Very few studies interpreted the result as indicating a lack of evidence against the null hypothesis of zero difference between the trial arms.

Strengths and limitations of this study

- We surveyed every issue of four journals for a recent 12-month period, hence the results are comprehensive and up to date.
- This was not a systematic review, but was restricted to four high-impact general journals. This means that we cannot draw any conclusions about other publications.
- Our classification system was developed by the authors and is not a validated tool.
- We only looked at reporting in abstracts; in the main text of papers authors may have made different and more accurate statements.

Introduction

Reports of randomised controlled trials (RCTs) usually attempt to draw conclusions about treatment effectiveness from their statistical analysis. It is common for results that pass a threshold for statistical significance, usually a p-value of less than 0.05, to be interpreted as indicating a real and clinically important effect. "Non-significance" (p>0.05) is often taken to mean that there is no difference between the treatments, or that the intervention is not effective. As has been pointed out many times, this is an erroneous conclusion.[1][2] Failure to reach a conventional threshold for "statistical significance" does not mean that it is safe to conclude that there is no difference. Every statistical test has a Type II error rate, which is the probability of obtaining a non-significant result, if the null hypothesis is false i.e. that there really is a difference. Trials are often designed with a 20% Type II error rate (80% power), for a true treatment effect of a specified size. With such a design, even if the true treatment effect is exactly as assumed (and designs often assume unrealistically large treatment effects), non-significance would be expected 20% of the time, and a conclusion of no difference would be wrong. Moreover, common issues such as fewer recruits than expected, more variability, or a lower incidence of outcomes, will reduce power, and make non-significant results more likely, even if in reality there is a real and important treatment effect. There is no way of discriminating between non-significant results that derive from chance or lack of power, and those that derive from a true lack of treatment benefit, except by more research.

Misinterpretation of non-significant results in clinical trials may be particularly damaging, because trials provide high-quality evidence, and their results often determine clinical guidelines and practice. Erroneous conclusions of ineffectiveness may result in non-adoption or abandonment of treatments that could actually be beneficial, and the existence of an apparently “definitive” trial that concluded ineffectiveness is likely to discourage further research. This problem was identified over twenty years ago[3] (“absence of evidence is not evidence of absence”), and subsequent studies have documented its persistence.[4][5]

The motivation for this study was our observation that, despite these warnings, poor interpretations of non-conclusive trial results remain common, even in the most prestigious journals. Many trials where the main results are not statistically significant conclude that there is no difference between the treatments, the intervention did not improve outcomes, or that it was not effective, none of which is a justified interpretation.

We examined how results were described in the Abstracts of recent reports of RCTs where the primary outcome did not show a statistically significant difference between the treatment arms, published in four leading general medical journals

Methods

We hand searched issues of four journals (New England Journal of Medicine (NEJM), Journal of the American Medical Association (JAMA), The Lancet and British Medical Journal (BMJ)) published between June 2016 and June 2017. Papers were included if they were primary reports of RCTs that had results for their primary outcome that were not “statistically significant” i.e. did not reach a prespecified threshold p-value that was regarded as indicating a true effect. We excluded non-inferiority, equivalence, single armed, dose-finding, and pharmacokinetic trials, as they have different reporting issues, and those that used Bayesian statistical methods. We included trials with more than two arms, and trials with multiple primary comparisons, if no treatment differences were claimed.

We extracted information from the abstract of each report on the description (from the Results or Findings section) and interpretation (from the Conclusions or Interpretation section) of the trials results for the primary outcome or outcomes. We concentrated on the abstracts because these are the most frequently viewed parts of papers, so conclusions expressed here will have the most impact. We classified the descriptions into ten categories (Table 1). The classification was developed at the start of the project, by reviewing trial reports from the same journals that were published in January to May 2016, the period immediately before our study's eligibility window. The classification made a distinction between reporting that claimed a lack of directional effect (e.g. "no improvement") and reporting that did not include any directional information (e.g. "no difference"), as well as whether the claim was qualified by reference to statistical significance (e.g. "no significant difference") or something else (e.g. "no substantial difference"). We created additional categories during the study, for reports that used methods that did not fit into any of the predetermined categories; for example, statements such as "there was a lack of evidence for a difference," or "treatments were similar," We also recorded whether confidence intervals and p-values were presented, and whether the confidence interval, or uncertainty more generally, were referred to in the conclusions.

Data were extracted by both authors independently and discrepancies resolved by discussion, leading to consensus in all cases. The authors were not blinded to the journals and authorship of individual articles.

Results

We identified 351 trial reports, of which 257 were not eligible, leaving 94 eligible papers, which reported 120 treatment comparisons (Figure 1). Three journals published the majority of studies (JAMA 28, Lancet 26, NEJM 32, BMJ 8). Significance tests were presented for 94/120 (78.3%) comparisons (79/94 papers (84%)), and confidence intervals for 96/120 (80%) comparisons (79/94 papers (84%)).

In the results section (Figure 2), the commonest reporting style was to present the point estimate and confidence interval, without any interpretation (58/120; 48.3%). A substantial number also referred to lack of statistical significance (38/120; 31.7%) or stated that there was no difference (9/120; 7.5%) or no improvement (7/120; 5.8%).

In the conclusions (Figure 3), a substantial majority of comparisons were classified as stating that there was no treatment benefit (65/120; 54.2%). The main alternative approach was to re-state the lack of a statistically significant difference (16/120; 13.3%) or lack of statistically significant benefit (14/120; 11.7%).

Results for papers rather than comparisons were similar (Table 1).

A threshold of $p < 0.05$ for statistical significance was used in all but two studies, which used lower values (0.04 and 0.01), as part of a correction for multiple comparisons. Similarly, all except these two studies used 95% confidence intervals. Confidence intervals for the main treatment comparison were presented by 79/94 studies (84.0%). This was surprisingly low, given that they have been a required part of trial reporting in the CONSORT guidelines for many years. Those that did not present confidence intervals for the main comparison either presented confidence intervals for the difference of each randomized group from baseline, or used only p-values. Both of these are poor reporting practices. The proportion of trials presenting p-values was the same (79/94; 84%). Sixty-four studies presented both confidence intervals and p-values, 15 confidence intervals without p-values, and 15 only p-values. Very few trials (3/94; 3.2%) explicitly referred to the confidence interval or uncertainty around the treatment effect estimate when drawing conclusions.

Discussion

Main results

Over 50% of the studies interpreted a non-significant result inappropriately, as indicating that there was "no difference" or "no benefit" to the intervention. Lack of statistical significance does not mean that no difference exists; this is one of the most basic misinterpretations of significance testing.[1]

Many of the studies that concluded a lack of benefit had substantial uncertainty about the direction and size of the treatment effect. For example, one trial concluded that the incidence of the outcome was "not reduced" by the intervention, based on a risk ratio of 1.13 (95% confidence interval 0.63, 2.00).[6] The confidence interval indicates that both substantial reduction or substantial increase (risk ratios as low as 0.63 or as high as 2.00) are compatible with the data, so the conclusion of no reduction does not seem justified. Conversely, some trials concluded "no benefit" when the results were actually strongly in one direction. One trial that concluded that the intervention was "not found to be superior," with a hazard ratio of 0.89 and a 95% confidence interval of 0.78 to 1.01 [7].

The conclusion seems inadequate; the study did suggest benefit, but not strongly enough to meet the arbitrary criterion for statistical significance. It does not seem reasonable for the conclusions from these two examples to be so similar, when the results are substantially different.

A further 24.3% of comparisons qualified their conclusion of lack of treatment benefit by referring to statistical significance (categories 2 and 5). This description is uninformative, because simply knowing that an arbitrary threshold was not achieved does not give much useful information, and relies on the reader being able to decode correctly what "significant" means in this context. It invites confusion between the technical meaning of "statistical significance" and the common English meaning of the word (important, substantial, worthy of attention), especially as results are often reported using phrases such as "not significantly different" or "no significant benefit" which can be read (and make sense) either as a statement about a formal statistical significance test, or as a regular English sentence. There is substantial empirical evidence that statistical significance is often misinterpreted by the public[8], academic researchers[9], and statisticians.[10]

Statements that the interventions were "similar" (category 8), there was no "substantial" difference (category 3) or no "clinically important" difference (category 6), which were used by smaller numbers of studies, are also difficult to interpret. None of them can be generally recommended as a way to describe non-significant results, but all might be appropriate in difference circumstances.

The most reasonable way to describe non-significant results is probably that the study did not find convincing evidence against the hypothesis that that the treatment effect was zero. Only one study contained a statement that referred to lack of evidence for a difference: "We found no evidence that an intervention comprising cleaner burning biomass-fuelled cookstoves reduced the risk of pneumonia in young children in rural Malawi," [11] describing an estimated incidence rate ratio of 1.01, with 95% confidence interval 0.91 to 1.13. Hence the data were compatible with either a small increase, or a small decrease, in the risk of pneumonia.

Statistical methods

All of the trials in our sample used traditional frequentist statistical methods. Although this is the dominant statistical methodology in clinical trials, there are many problems in the understanding and interpretation of p-values, significance tests,[12][13][1] and confidence intervals,[14][15] which have recently received substantial publicity, in the wake of publication of the American Statistical Association's guidance on p-values and significance testing[2] and more recent publications[16-19].

One important issue is the use of a threshold for "significance," creating a binary classification of results, which is usually interpreted as indicating treatments that "work" and "don't work" (or "positive" and "negative" trials, or "effective" and "ineffective" treatments).[19-21] In reality there is no such sharp dividing line between treatments that work and do not work, and significance tests simply impose an arbitrary criterion. The persistence of dichotomisation of results may be largely due to an unrealistic expectation that trials will provide certainty in their conclusions and treatment recommendations. Sometimes trials will reduce our uncertainty sufficiently that the best clinical course of action is clear, but often they will not. An argument that is often advanced in favour of dichotomization of results is that because many trials seek to inform clinical practice, a decision needs to be made about whether the intervention should be used in patient care. The counter-argument to this is that decisions about use of healthcare interventions should be based not on whether or not a single primary outcome reaches an arbitrary significance threshold, but on consideration of the overall benefits, harms and costs of the intervention, using appropriate decision modelling methodology.

Improving the language for describing results

One straightforward way to improve reporting of results is to be more careful about the language that is used to describe them and draw conclusions, and ensure that written descriptions match the numerical results. We should avoid language that is ambiguous or open to misinterpretation, for example only describing treatments as ineffective if we have a high degree of confidence that the treatment does not have clinically important effects. We should also pay more attention to uncertainty, and consider what possible values of the unknown underlying treatment effect could have given rise to the data that were observed. Often, this range will be wide. We should not expect every trial to lead to a clear treatment recommendation, but be honest about the degree to which a study is able to reduce our uncertainty. Confidence intervals were originally promoted for trial reporting to encourage this sort of interpretation, and to avoid the false certainty provided by significance tests.[22-23] But even though most trials now present them, they are rarely considered in the conclusions,[24-25] and are often used simply as an alternative way to perform significance tests, concentrating only on whether the confidence interval excludes the null value.

A recent online discussion [26] about language for describing frequentist trial results gave some examples of accurate statements that could be used. Three examples of statements for trials that did not find a treatment difference, from this discussion, are given in Table 2. These statements are very different from those used by most of the papers in our sample, and make much more limited claims than many real papers. However, these claims accurately reflect the conclusions that can be drawn from frequentist statistical analyses. More accurate language would help to prevent common over-interpretations, such as the belief that non-significance means that a treatment difference of zero has been established.

Improving the statistical methods

A more radical solution is to change the statistical approach that we use. One fundamental problem with traditional frequentist statistical methods is that they do not provide the results that clinicians, policy makers and patients actually want to know: what are the most plausible values of the treatment effect, given the observed data? Significance tests actually do the reverse; they calculate probabilities of the data (or more extreme data), assuming a specific null value of the treatment effect. This is a major reason why reporting frequentist results accurately is so convoluted, and why they are so difficult to understand. However, easily-interpretable probabilities of clinically relevant results can be readily obtained using Bayesian methods. The output from a Bayesian analysis is a probability distribution giving, the probability of all possible values of the treatment effect, taking into account the trial's data, and usually (via the prior), external information as well. We can use this distribution (the posterior probability distribution) to calculate relevant and informative results, such as the probability of a benefit exceeding a threshold for clinical importance, the probability of the treatment effect being within a range of clinical equivalence, or the range of treatment effects with 95% probability (or 50%, or any other value). Some examples of the sorts of informative statements that can be made from Bayesian results are given in a blog post by Frank Harrell [27]. A particular advantage is that, with Bayesian methods, there is no need to reduce results to a dichotomy, but instead we can refer directly to probabilities of events of interest.

Limitations of the study

This study looked only at reporting of results in abstracts of published RCTs. We concentrated on abstracts because they are the most frequently read parts of papers, and always report the main results. They are therefore likely to be particularly important in determining readers' interpretation of the trial's results. It is possible that in other parts of the papers, reporting may have been different, and potentially more accurate. However, this is much harder to assess because results are typically reported in several different places, and often inconsistently.

We concentrated on four of the highest profile general medical journals. Obviously, RCTs are also published in a large number of other, more specialised, journals, but we cannot say whether they have the same issues of reporting as we found. Our expectation would be that, as the journals we selected are seen as some of the most prestigious publications, reporting problems would be at least as common elsewhere.

Our classification of reporting types was invented by the authors, and is not intended as a general tool for conducting this type of study. However, we feel that it is a reasonable classification that makes distinctions between the different types of reporting that we wished to identify.

Conclusions

Despite many years of warnings, inappropriate interpretations of RCT results are widespread in the most prestigious medical journals. We speculatively suggest several possible factors that may be responsible. First, authors and editors may want to present a clear message, and there is a widespread expectation that RCTs should result in clear recommendations for clinical practice. It is easier to understand a conclusion that "X did not work" than a complicated statement that more accurately reflects what a non-significant result means. Second, use of significance testing as the main analytical method provides a ready means of dichotomization of results, encouraging an over-simplified binary interpretation of interventions. Third, the general difficulty of understanding frequentist results means that correct interpretation is convoluted and difficult to relate to real life.

We suggest that interpretation of results should pay more attention to uncertainty and the range of treatment effects that could plausibly have given rise to the observed data. Use of Bayesian statistical methods would facilitate this by addressing the clinical questions of interest directly.

Funding and Ethics

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. Ethics Committee approval was not required.

Copyright Statement

Competing Interests

Both authors have completed the Unified Competing Interest form and declare no support from any organisation, no financial relationships with any organisations that might have an interest in this work, and no other relationships or activities that could appear to have influenced the submitted work.

Contributors

Simon Gates designed the study, assisted with data extraction, performed the analysis and drafted the manuscript. Elizabeth Ealing collected the data, assisted with analysis, and revised the manuscript. Simon Gates is the guarantor.

Transparency statement

The lead author affirms that this manuscript is an honest, accurate, and transparent account of the study being reported.

Patient and public involvement

There was no patient and public involvement in this study.

Data Sharing Statement

The data from this study are available from Open Science Framework (https://osf.io/chsva/files/)

References

[1] Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., et al. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. European journal of epidemi- ology 2016;31(4):337–350.

[2] Wasserstein, R., Lazar, N. The asas statement on p-values: context, process, and purpose. The American Statistician 2016;70(2):129–133.

[3] Altman, D.G., Bland, J.M. Statistics notes: Absence of evidence is not evidence of absence. BMJ 1995;311(7003):485.

[4] Alderson, P., Chalmers, I. Survey of claims of no effect in abstracts of Cochrane reviews. BMJ 2003;326(7387):475.

[5] Greenland, S. Null misinterpretation in statistical testing and its impact on health risk assessment. Preventive medicine 2011;53(4):225–228.

[6] Thomusch, O., Wiesener, M., Opgenoorth, M., Pascher, A., Woitas, R.P., Witzke, O., et al. Rabbit-atg or basiliximab induction for rapid steroid withdrawal after renal transplantation (harmony): an open-label, multicentre, randomised controlled trial. The Lancet 2016;388(10063):3006–3016.

[7] Johnston, S.C., Amarenco, P., Albers, G.W., Denison, H., Easton, J.D., Evans, S.R., et al. Ticagrelor versus aspirin in acute stroke or transient ischemic attack. New England Journal of Medicine 2016;375(1):35– 43.

[8] Tromovitch, P. The lay public's misinterpretation of the meaning of 'significant': A call for simple yet significant changes in scientific reporting. Journal of Research Practice 2015;11(1):1.

[9] Haller, H., Krauss, S. Misinterpretations of significance: A problem students share with their teachers. Methods of Psychological Research 2002;7(1):1–20.

[10] McShane, B.B., Gal, D. Blinding us to the obvious? the effect of statistical training on the evaluation of evidence. Management Science 2015;62(6):1707–1718.

[11] Mortimer, K., Ndamala, C.B., Naunje, A.W., Malava, J., Katundu, C., Weston, W., et al. A cleaner burning biomass-fuelled cookstove intervention to prevent pneumonia in children under 5 years old in rural Malawi (the Cooking and Pneumonia Study): a cluster randomised controlled trial

[12] Goodman, S. A dirty dozen: twelve p-value misconceptions. In: Seminars in hematology; vol. 45. Elsevier; 2008, p. 135–140.

[13] Goodman, S.N. Toward evidence-based medical statistics. 1: The p value fallacy. Annals of internal medicine 1999;130(12):995–1004.

[14] Hoekstra, R., Morey, R.D., Rouder, J.N., Wagenmakers, E.J. Robust misinterpretation of confidence intervals. Psychonomic bulletin & review 2014;21(5):1157–1164.

[15] Morey, R.D., Hoekstra, R., Rouder, J.N., Lee, M.D., Wagenmakers, E.J. The fallacy of placing confidence in confidence intervals. Psychonomic bulletin & review 2016;23(1):103–123.

[16] Amrhein, V., Greenland, S., McShane, B.. Scientists rise up against statistical significance. Nature 2019;567(7748):305–307

[17] Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nat Hum Behav. 2018 Jan;2(1):6-10

[18] Lakens D, Adolfi FG, Albers CJ, Anvari F, Apps MAJ, Argamon SE, et al. Justify your alpha. Nat Hum Behav 2018; 2, 168–171

[19] McShane, B.B., Gal, D. Statistical significance and the dichotomization of evidence. Journal of the American Statistical Association 2017;112(519):885–895.

[20] Senn, S. Dichotomania: an obsessive compulsive disorder that is badly affecting the quality of analysis of pharmaceutical trials. Proceedings of the International Statistical Institute, 55th Session, Sydney 2005.

[21] Gelman, A., Stern, H. The difference between significant and not significant is not itself statistically significant. The American Statistician 2006;60(4):328–331.

[22] Moher, D., Hopewell, S., Schulz, K.F., Montori, V., Gøtzsche, P.C., Devereaux, P., et al. Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869.

[23] Gardner, M.J., Altman, D.G. Confidence intervals rather than p values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed) 1986;292(6522):746–750.

[24] Fidler, F., Thomason, N., Cumming, G., Finch, S., Leeman, J. Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. Psychological Science 2004;15(2):119–126.

[25] Gewandter, J.S., McDermott, M.P., Kitt, R.A., Chaudari, J., Koch, J.G., Evans, S.R., et al.

Interpretation of cis in clinical trials with non-significant results: systematic review and recommendations. BMJ Open 2017;7(7):e017288.

[26] https://discourse.datamethods.org/t/language-for-communicating-frequentist-results-about-treatment-effects/934

[27] http://www.fharrell.com/post/bayes-freq-stmts/

Table 1. Categories of reporting of RCT results in Results and Conclusion sections of abstracts. For trials with multiple results, all were reported in the same way in all trials except one (ref); for this trial, we have included the results for the survival co-primary outcome rather than the ordinal composite outcome.

Category	Description	Examples	Number of comparisons (%) (n=120)		Number of papers (%) (n=94)	
			Results	Conclusions	Results	Conclusions
1	Statement of no difference between treatments	"did not differ," "no difference," "no effect," "no change."	8 (7.2)	5 (4.5)	6 (7.1)	1 (1.2)
2	Statement that there was no difference between treatment, qualified by reference to statistical significance	"no significant difference," "not statistically different," "not statistically significant," "no significant effect."	34 (30.6)	13 (11.7)	25 (29.4)	12 (14.1)
3	Statement that there was no difference between treatments, qualified by something other than statistical significance	"no substantial difference," "no clinically relevant difference."	3 (2.7)	1 (0.9)	2 (2.4)	1 (1.2)
4	Statement that the intervention was not beneficial	"did not result in increase (or decrease or improve)," "was not superior," "did not increase (or decrease or improve)," "did not prevent."	7 (6.3)	61 (55.0)	3 (3.5)	46 (54.1)
5	Statement that the intervention was not beneficial, qualified by reference to statistical significance	"not significantly better (or worse)," "did not significantly increase (or decrease)," "not statistically increased (or decreased)."	1 (0.9)	14 (12.6)	1 (1.2)	12 (14.1)
6	Statement that the intervention was not beneficial, qualified by	"not substantially increased (or decreased)."	0 (0)	3 (2.7)	0 (0)	2 (2.4)

	reference to something other than statistical significance					
7	Statement that there was a lack of evidence for a difference.	“no evidence that [intervention] reduced the risk of [outcome]”	0 (0)	1 (0.9)	0 (0)	1 (1.2)
8	Statement that the treatments compared were similar.	“yield similar outcomes” “similar risk of [outcome]” “rate of [outcome] was similar”	3 (2.7)	7 (6.3)	3 (3.5)	4 (4.7)
9	Quotation of the results, without any claim about the size or direction of effect.	Estimate and 95% confidence interval	55 (49.5)	4 (3.6)	45 (52.9)	4 (4.7)
10	Clinical recommendation, without interpretation of results.	“There is no harm in [using intervention]” “The choice between [interventions] should be made based on clinical knowledge”	0 (0)	2 (1.8)	0 (0)	2 (2.4)

Table 2. Examples of accurate statements for describing non-significant frequentist results, from https://discourse.datamethods.org/t/language-for-communicating-frequentist-results-about-treatment-effects/934 [23], concerning a hypothetical trial that evaluating differences in systolic blood pressure (SBP).

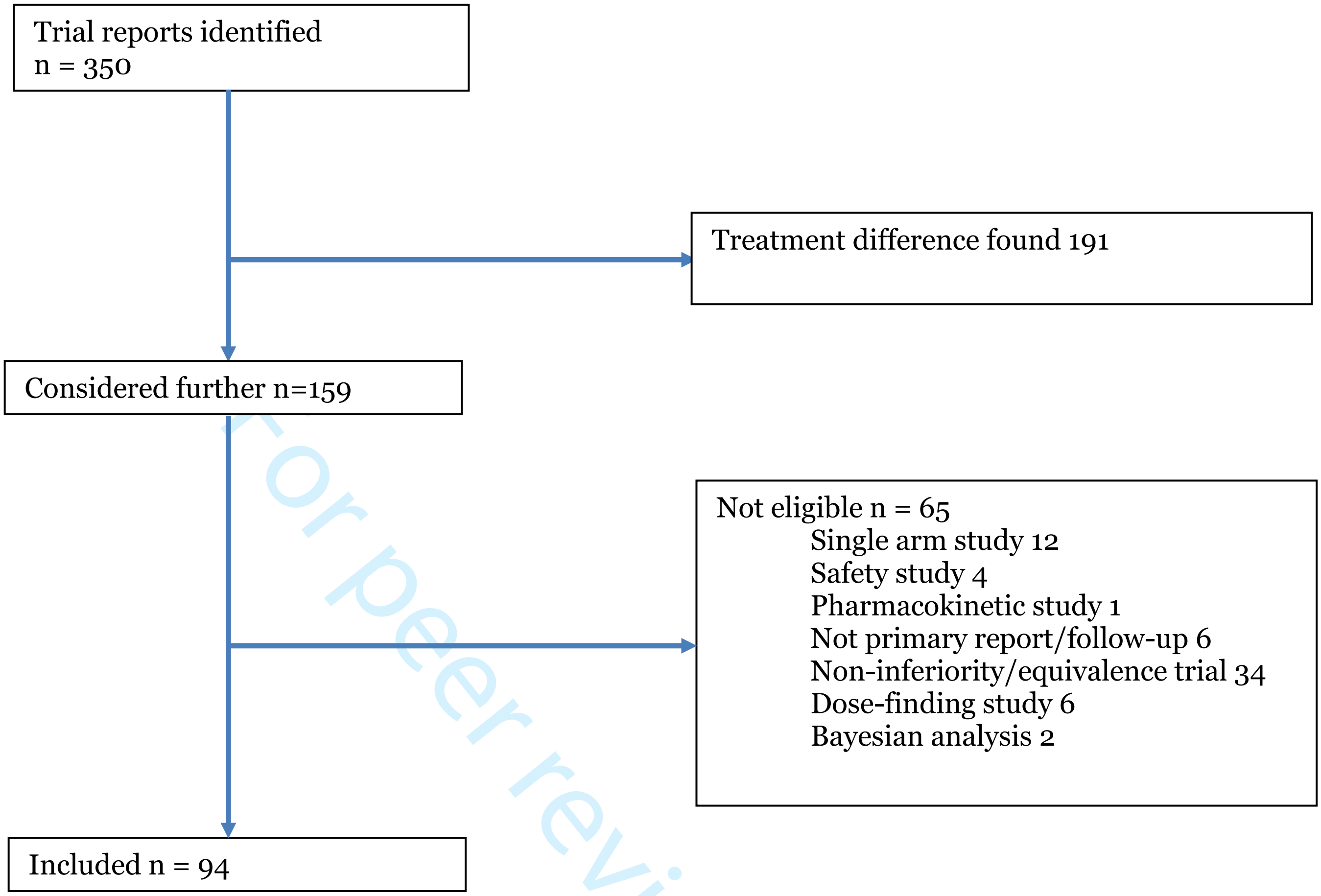
Example 1	We were unable to find evidence against the hypothesis that A=B (p=0.4) with the current sample size. More data will be needed. As the statistical analysis plan specified a frequentist approach, the study did not provide evidence of similarity of A and B.
Example 2	Assuming the study's experimental design and sampling scheme, the probability is 0.4 that another study would yield a test statistic for comparing two means that is more impressive that what we observed in our study, if treatment B had exactly the same true mean SBP as treatment A.
Example 3	Treatment B was observed in our sample of n subjects to have a 4mmHg lower mean SBP (systolic blood pressure) than treatment A with a 0.95 2-sided compatibility interval of [-13, 5], indicating a wide range of plausible true treatment effects. The degree of evidence against the null hypothesis that the treatments are interchangeable is p=0.11.

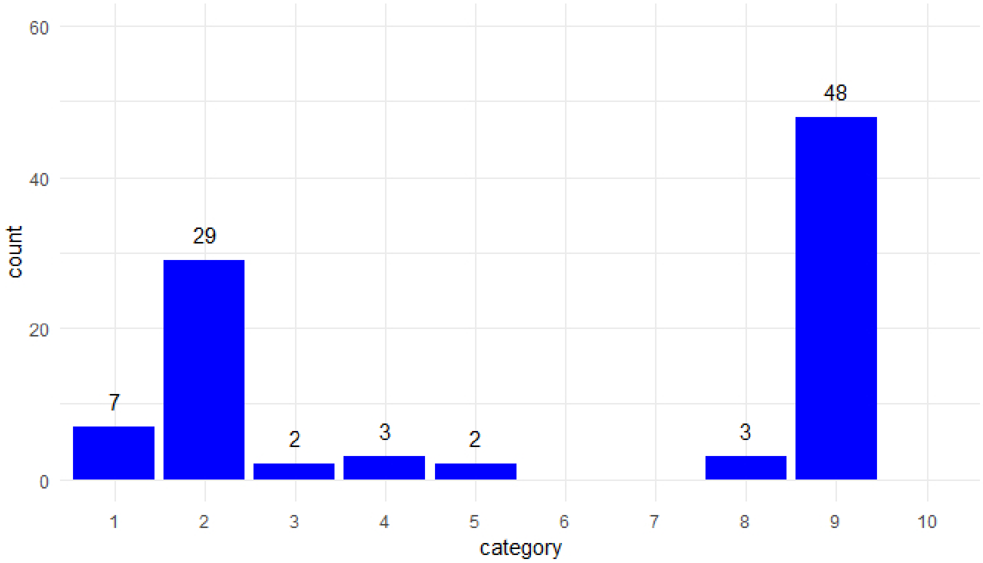
Figures

Figure 1: Flowchart of studies.

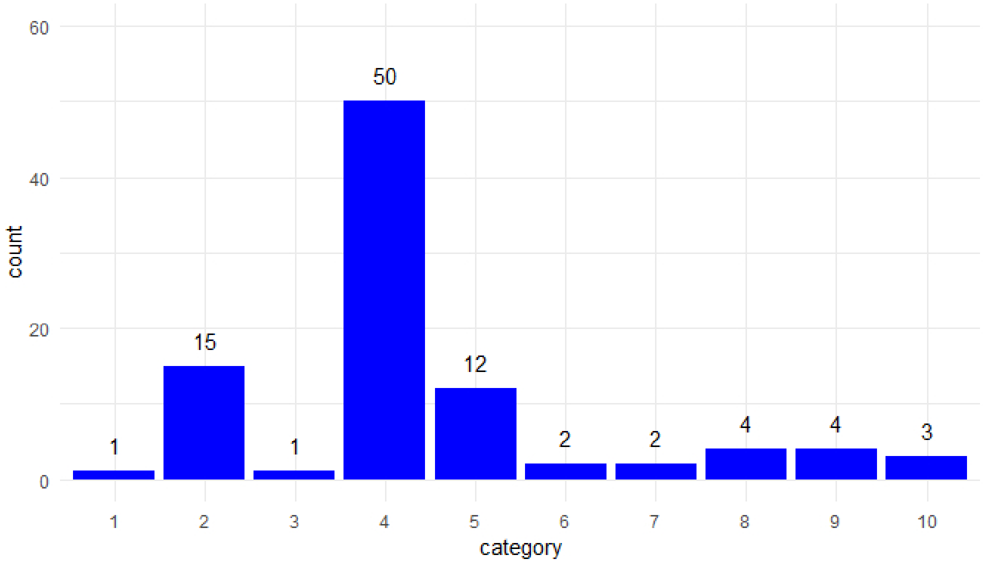
Figure 2: Frequencies of different types of description of results in Results section of abstracts. Categories (described fully in Table 1): 1. no difference; 2. no statistically significant difference; 3. no substantial or clinically important difference; 4. no improvement or no treatment benefit; 5. no significant improvement; 6. no substantial improvement; 7. lack of evidence for a difference; 8. treatments were similar; 9. statement of results; 10. clinical recommendation

Figure 3: Frequencies of different types of description of results in Conclusions section of abstracts. Categories (described fully in Table 1):): 1. no difference; 2. no statistically significant difference; 3. no substantial or clinically important difference; 4. no improvement or no treatment benefit; 5. no significant improvement; 6. no substantial improvement; 7. lack of evidence for a difference; 8. treatments were similar; 9. statement of results; 10. clinical recommendation.


Trial reports identified
n = 350
Treatment difference found 191
Considered further n=159
Not eligible n = 65
Single arm study 12
Safety study 4
Pharmacokinetic study 1
Not primary report/follow-up 6
Non-inferiority/equivalence trial 34
Dose-finding study 6
Bayesian analysis 2
Included n = 94



Frequencies of different types of description of results in Results section of abstracts. Categories (described fully in Table 1): 1. no difference; 2. no statistically significant difference; 3. no substantial or clinically important difference; 4. no improvement or no treatment benefit; 5. no significant improvement; 6. no substantial improvement; 7. lack of evidence for a difference; 8. treatments were similar; 9. statement of results; 10. clinical recommendation



Frequencies of different types of description of results in Conclusions section of abstracts. Categories (described fully in Table 1):): 1. no difference; 2. no statistically significant difference; 3. no substantial or clinically important difference; 4. no improvement or no treatment benefit; 5. no significant improvement; 6. no substantial improvement; 7. lack of evidence for a difference; 8. treatments were similar; 9. statement of results; 10. clinical recommendation.